DOCUMENT RESUME

ED 178 556 TM 009 545

AUTHOR Benson, Jeri
TITLE A Comparison of the One-and Three-Parameter Logistic Models on Measures of Test Efficiency.
PUB DATE Apr 79
NOTE 27p.; Paper presented at the annual meeting of the National Council on Measurement in Education (San Francisco, CA, April 9-11, 1979)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Statistics; Computer Programs; Costs; *Efficiency; *Item Analysis; Mathematical Models; Senior High Schools; Statistical Analysis; *Test Construction; Test Items; Test Reliability
IDENTIFIERS *Rasch Model; *Three Parameter Model

ABSTRACT

Two methods of item selection were used to select sets of 40 items from a 50-item verbal analogies test, and the resulting item sets were compared for relative efficiency. The BICAL program was used to select the 40 items having the best mean square fit to the one parameter logistic (Rasch) model. The LOGIST program was used to select the 40 items with the highest discrimination parameter. Selection was based on two independent samples of 2,500 randomly assigned from a group of 5,000. The set of items selected by the three-parameter program was somewhat more difficult than the set selected by the one-parameter program, but 31 items were common to both sets and there were no other significant differences. The relative efficiency of the two sets of items showed no real differences. It was concluded that both programs tended to select the better items from a well developed item pool, but that the LOGIST program required considerably more computer time. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED178556

A Comparison of the One- and Three-Parameter Logistic Models on Measures of Test Efficiency

Jeri Benson

University of Southern California

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Jeri Benson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A paper presented at the Annual meeting of the National Council on Measurement in Education, San Francisco, California, April 1979.

TM 009 545

Abstract

The purpose of this study was to empirically compare tests developed by the one- and three-parameter logistic models in terms of relative efficiency. The data employed were a 50 item verbal analogy test from a sample of 5000 high school students. The procedures included item selection, computation of ability estimates based on cross-validation samples, and a relative efficiency comparison of the two tests. The results indicated that cognitive tests developed using the one- and three-parameter logistic models did not differ in terms of their relative efficiency for most ability groups.

A Comparison of the One- and Three-Parameter Logistic Models on Measures of Test Efficiency[1]

The concept of latent trait theory dates back to Lawley (1943) and Lazarsfeld (1950); however, it is Frederic Lord who is generally given credit as the father of latent trait theory (1952, 1953). The application of latent trait methodology to the area of education measurement became a practical reality with the publication of Lord and Novick's Statistical Theories of Mental Test Scores (1968) and the availability of computer programs. Moreover in the last five years increased interest in the application of latent trait theory to testing problems has been observed by the large number of papers published on the topic, most notably, the entire issue of the Journal of Educational Measurement, 14(2), 1977 was devoted to latent trait methodology.

Latent trait theory specifies a relationship between the observable examinee test performance and the unobservable traits or abilities assumed to underlie performance on a test (Hambleton & Cook, 1977). The relationship is described by a mathematical function; hence latent trait models are mathematical models. Several latent trait models have been developed for use with dichotomously scored data: the normal ogive, and the one-, two-, and three-parameter logistic models. These models along with their mathematical formulations have been described by Hambleton and Cook (1977) and Lord and Novick (1968). All four models are based on the assumption that the items in the test are measuring one common

[1]Efficiency has been described by Lord (1974a, 1974b) and is used as a measure of the effectiveness of one test over another as a function of ability level.

ability and the assumption of local independence exists between the items and the examinees. These two assumptions imply that a test which measures only one trait or ability will have less measurement error in the test score than a test that is multidimensional, and that the response of an examinee to one item is not related to his response to any other item. Where the latent trait models begin to differ is with respect to the shape of their item characteristic curves.

The normal ogive, developed by Lord (1952, 1953), and the two-parameter logistic model, developed by Birnbaum (1968) are essentially the same function (Baker, 1961), each providing an estimate of the item difficulty and discrimination parameters. The three-parameter model also developed by Birnbaum (1968), is an extension of the two-parameter model where the additional parameter being estimated is an index for guessing. The one-parameter logistic model, developed by Rasch (1960), is commonly referred to as the Rasch model. In addition to the two assumptions of latent trait models mentioned above, the Rasch model makes a third assumption, that all item discriminations are equal. Thus, the only parameter to be estimated using the Rasch model is item difficulty.

A review of the literature in the area of latent trait theory yielded several comparative studies involving latent trait models. One of the early studies of Birnbaum's three-parameter model was conducted by Lord (1968). The study focused on developing tailored tests via item and test information curves using data from the Scholastic Aptitude Test (SAT) verbal subtest. While this was not a comparison among latent trait models,

it is important as it represents one of the first applications of latent trait theory to practical testing problems. One of the first comparative studies was conducted by Hambleton and Traub (1971). They compared the efficiency of ability estimates provided by the one- and two-parameter model to the three-parameter model using Birnbaum's concept of information (1968). Eleven simulated tests of 15 items each were generated varying item discrimination and degree of guessing. The authors sought to determine how efficient the one- and two-parameter logistic models were under these conditions taking the three-parameter model to be the true model. The results indicated that when guessing was a factor the three parameter model was most efficient in providing ability estimates, but when guessing was not a factor all models were equally efficient.

In a later study by Hambleton and Traub (1973), the one- and two-parameter models were compared with three sets of real data (the verbal and mathematics subtests of the Ontario Scholastic Aptitude Test, and the verbal subtest of the SAT, items = 45, 20, and 80 respectively). Their results indicated that generally the two-parameter model fit the data better than the one-parameter model. The loss in predicting performance was greatest for shorter tests and smallest for longer tests. These findings confirm Birnbaum's conjecture (1968, p. 492) that if the number of items in a test is very large the inferences that can be made about an examinee's ability will be much the same whether the one- or the two-parameter logistic model is used. The authors questioned whether the gain obtained with the two-parameter model is worth the increased

computer time of estimating the item discrimination parameter.

Reckase (1977) compared the one- and three-parameter models using achievement test data. His results indicated that the one-parameter model provided ability estimates that were equivalent to the estimates provided by the three-parameter model when unifactor tests were used. The three-parameter model gave better item data fit than the one-parameter model. However, the superiority of the three-parameter model on measures of item data fit were at the expense of greater computer costs. His conclusion, as well as the conclusion made by Hambleton and Traub (1973), was whether the increased computer cost for the three-parameter model was worth the improvement obtained from using a more complex model. Based on these three comparative studies, it was concluded that more precise item data fit was achieved when the three-parameter model was used. However, when the tests being compared were unidimensional, the observed ability estimates were similar using either the one- or three-parameter model.

In test construction, the precision of ability estimates is a crucial question. Thus, the superiority of one latent trait model over another may be viewed in terms of test precision. However, a single overall measure of test precision, e.g., reliability estimates is very limited. It would be more useful to have a measure of a test's precision at various points along the ability continuum. This is the concept of test information, and the ratio of two test information functions is defined as their relative efficiency (Birnbaum, 1968, p. 472). Two studies have

been noted where relative efficiency comparisons were made between tests developed through classical test procedures and latent trait procedures (Benson, Crocker, & Ware, 1978; Benson & Wilson, 1979). However, few studies have been conducted where tests developed by various latent trait methods were compared in terms of their relative efficiency. Therefore, the purpose of this study was to empirically compare the relative efficiency of tests developed using two latent trait methods, e.g., the one- and three-parameter logistic models. To ensure that the same trait was being measured by both tests a common item population was used. The study differed from Hambleton and Traub's 1971 study in that real data rather than simulated data were used and test length was longer and more realistic.

Method

The Sample

A population of 78,751 was available from 435 high schools throughout the state. From this population a 1 in 15 systematic sample of students was selected (Mendenhall, Ott, & Scheaffer, 1971) to obtain a sample of 5000 students. The sample of 5000 was then randomly assigned to one of the two test development techniques, 2500 in each group.

The Instrument

The instrument selected for use in this study was the Verbal Aptitude subtest of a statewide assessment battery, developed by the Educational Testing Service. The Verbal Aptitude subtest is comprised of 50 verbal

analogies in a multiple choice format. One use of the scores is for college admission decisions. The instrument was selected because it is highly representative of other national aptitude tests used for college admissions, and this subtest has been shown to be unidimensional, e.g., the items measure a single ability (Benson, Crocker, & Ware, 1978). Thus the assumption of unidimensionality for both latent trait models was met.

The Procedure

The data were analyzed in three phases. First, the 40 items best fitting the one- and three-parameter models were selected. The BICAL program (Wright & Mead, 1976) was used to select items for the one-parameter model. The mean square fit of the item to the model was the criterion for item selection. The LOGIST program (Wood, Wingersky, & Lord, 1976) was used to select items for the three-parameter model. The items with the highest discrimination parameter (a_i) was the criterion for item selection. Second, a cross-validation technique was used to obtain descriptive statistics for the two 40 item tests, as well as student ability estimates. Unweighted item scores were used to obtain the ability estimates in both models. Third, a comparison of the relative efficiency of the two 40 item tests was made using the ability estimates and procedures described by Lord (1974a, 1974b).

Relative efficiency is a procedure to compare two unidimensional tests that measure the same trait. In classical test theory, two tests that measure the same ability are usually compared in terms of their

reliability coefficients, but this only gives a single overall comparison. The formula developed by Lord for relative efficiency provides a more precise way of comparing two tests that may not be of equal difficulty, but still measure the same ability. The formula for approximating relative efficiency is (Lord, 1974b, p. 248):

$$\text{R.E. } (y,x) = \frac{n_y}{n_x} \cdot \frac{x(n_x - x)f_x^2}{y(n_y - y)f_y^2},$$

where R.E. (y,x) denotes the relative efficiency of y compared to x, n_y and n_x denote the number of items on the two tests, x and y are the number-right scores having the same percentile rank, an f_x^2 and f_y^2 are the squared observed frequencies of x and y. Lord has suggested that this formula only be used when the sample size is large and the tests are not extremely short. Relative efficiency estimates using the above formula provide information on both tests at varying points along the ability continuum.

Results

Prior to item selection, the equivalency of the two random samples was determined by obtaining basic descriptive statistics for the 50 item test for the population and each sample. This information is reported in Table 1.

Insert Table 1 About Here

As noted on Table 1, sample one was assigned to the one-parameter logistic model (BICAL procedure) and sample two was assigned to the three-parameter logistic model (LOGIST procedure). The differences between the sample means and the population mean were not statistically significant. Therefore it was assumed that there were no systematic differences between the two samples before the items were selected.

The BICAL procedure was used to select the best 40 items fitting the one-parameter model. The mean square fit statistic (Wright & Panchapakesan, 1969), a measure of item fit to the Rasch model, was the criterion for item selection. The closer the mean square fit was to zero the better the item fit the model, thus the 40 items with the lowest mean square fit statistic were selected.

The LOGIST procedure was used to select the best 40 items fitting the three-parameter model. The criterion for item selection was based on item discrimination parameters (a_i). Thus, the 40 items with the highest a_i values were selected as the items best fitting the three-parameter model.

Descriptive statistics for the items selected by each latent trait model are listed in Table 2.

Insert Table 2 About Here

Thirty one of the forty items (78 percent) were common to both procedures, and nine items were unique to each procedure (22 percent). One item (# 14) was not selected by either procedure. Generally, the more difficult items were selected by the three-parameter model. This finding was also noted on Tables 3 and 4, where the item parameter estimates and the descriptive data for the two 40 item tests, respectively, were presented.

Insert Tables 3 and 4 About Here

As noted on Table 3, the item difficulty estimate for the one-parameter model has a mean equal to .01 and a standard deviation equal to .98. These estimates are very close to the expected values of zero and one for the item difficulty mean and standard deviation (Wright & Mead, 1975) indicating the degree fit of the 40 items to the model. Item difficulty values can range between $\pm$ three for the one-parameter model. The difficulty estimates ranged from -2.36 to 1.85, a slight tendency to select easier items was noted. The one-parameter model does not provide a parameter for item discrimination as all item discriminations are considered equal and centered at one (Wright & Mead, 1975).

The item parameter estimates for the three-parameter model are also reported on Table 3. Item discrimination for the three-parameter model can range between zero and two. The estimates for the 40 item test ranged from .81 to 2.00 with four items reaching a value of 2.00. The

mean item discrimination was 1.26, slightly higher than the assumed value of one for the one-parameter model. The item difficulty estimates for the three-parameter model, like in the one-parameter model range, between $\pm$ three. The actual difficulty estimates for the three-parameter model ranged from -2.17 to 2.13, a few more difficult items were selected using the three-parameter model. The guessing parameter estimates for the items ranged from .11 to .43 with a mean value of .22, which is very close to the .25 expected value for a four option multiple choice test.

The data reported in Table 4 indicated the means of the two tests differed significantly ($p < .05$), the one-parameter model produced an easier test as noted by the mean test score of 20.53 compared to the mean test score for the three-parameter model of 18.64. The differences represented a confirmation of the item statistics data from Table 3. The standard error of measurement for the one-parameter model was less than the standard error of measurement for the three-parameter model by a difference of .31 points (SEM = 2.51, 2.82 respectively). This difference was primarily a function of the greater variability in the test produced by the three-parameter model.

Both latent trait techniques produced tests with similar internal consistency estimates. However, this single overall measure of a test's precision does not provide enough information to judge the effectiveness of one model over another in terms of the precision of measurement along the ability continuum. Therefore, the relative efficiency of the three-parameter model was compared to the one-parameter model, and the results

are shown in Figure 1.

Insert Figure 1 About Here

The relative efficiency was plotted along the ordinate, while the percentile rank (ability level) was plotted along the abscissa. Since the tests were composed of the same number of items, a relative efficiency of 1.00 would indicate that the tests were equally efficient. Computed values for the relative efficiency comparisons have been reported in Appendix A.

Generally the data displayed in Figure 1 indicated no real differences in the relative efficiency of the two tests, as noted by the spread of the points around 1.0. There were two exceptions, however, at the 30th and 99th percentile ranks, where the test based on three-parameter model was more efficient than the test based on the one-parameter model.

Discussion and Conclusions

From the data reported on Tables 3 and 4 it was found that the test based on the three-parameter model contained items that were slightly more discriminating ($\overline{X}$ = 1.26) and more difficult ($\overline{X}$ = .74) than the test based on the one-parameter model ($\overline{X}$ = 1 and $\overline{X}$ = .01 for item discrimination and difficulty respectively). This finding was reflected in the mean test scores for the two 40 item tests. The mean for the test based on the three-parameter model containing the more discriminating and difficult items was 18.64. The mean for the test based on the one-para-

meter model was 20.53. The difference was found to be significant at the .05 level. Despite the significant finding the relative efficiency of the two tests was essentially the same as illustrated in Figure 1.

A careful examination of the formula for relative efficiency, reported on page 6, indicated that when $n_x = n_y$ and $x = y$, that it was the number of examinees at a specified ability level (f_x^2 and f_y^2) that determined the relative efficiency of a test. That is, the fewer examinees observed at a particular percentile rank, the better the test discriminates at that percentile rank. Therefore, test efficiency was equated with the level of discrimination the test was able to make between examinees, at various scores or percentile ranks. Thus, even though the two latent trait procedures produced different tests in terms of item and test statistics their relative efficiency estimates were very similar. One fact that should not be overlooked was that of the 50 items in the population 31 were common to both procedures. Based on the results of this study, it was concluded that both models tended to select the better items (from a well developed item population), even though item difficulty and discrimination estimates for the two tests were slightly different, hence the similarity in test efficiency.

However, a major mathematical difference between the one- and three-parameter logist models is the estimation of two additional item parameters. In the three-parameter model the item response is weighted in order to estimate the ability, the item discrimination, and the guessing parameters. The weighting is an iterative process that may never converge or stabilize unless arbitrary boundaries are established (Wright, 1977).

In this study, 65 ability estimates out of 2500 (3 percent) failed to converge, and were arbitrarily set at - 4.00. Because of the estimation procedure, the three-parameter logistic model is much more costly in terms of computer time than the one-parameter logistic model. For this study, the three-parameter model took 18 minutes and 26 seconds of CPU time using the sixteen iteration stages defined in the LOGIST program. The one-parameter model took 1 minute and 25 seconds of CPU time using the BICAL program. In addition, the sometimes lack of convergence in ability estimation from the three-parameter model leads one to believe the ability estimates obtained through the one-parameter model are more stable. In fact, only the one-parameter Rasch model provides a ratio scale of measurement in terms of the callibrated item and ability scores (Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1977).

If a test is to be constructed using latent trait methods to discriminate at certain ability levels as has been suggested by Lord (1968), the choice of which logistic model to use may make a difference. However, based upon the results of this study it was concluded that the simple one-parameter logistic model was equally as efficient as the more complex three-parameter model when a well defined item pool was used. This conclusion was based on the results of the relative efficiency comparisons, the fact that 78% of the items selected were common to each procedure, and the excessive cost of estimating the two additional item parameters. Thus, the results of this study indicate the simple one-

parameter model was preferred to the more complex three-parameter model in terms of test development. In addition, previous researchers (Hambleton & Traub, 1973; Reckase, 1977) using different data bases have made similar conclusions especially when the computer costs of the three-parameter model were considered. Before the findings of the present study can be generally accepted, replication is needed using not only other populations, but other instruments as well. It may be important to consider whether similar results would be obtained if another instrument were used whose items were not as well refined as the ones used in the present study.

Table 1

Descriptive Statistics for the 50 Item Subtest Based on the Population and Each Random Sample

	Population N = 78,751	Sample 1 (BICAL) N = 2500	Sample 2 (LOGIST) N = 2500
Mean	25.95	25.99	25.82
Standard Deviation	8.23	8.23	8.26
Internal Consistency	.88	.86	.86
Standard Error of Measurement	2.85	3.08	3.09

Table 2

Descriptive Statistics for Items Selected by Each Latent Trait Model[a]

Item	Difficulty	Discrimination[b]	Selected Item[c]
1	.74	.43	B
2	.90	.59	L, B
3	.78	.52	B
4	.81	.43	B
5	.84	.62	B
6	.80	.63	L, B
7	.61	.74	L
8	.74	.42	B
9	.67	.60	L, B
10	.75	.69	L
11	.70	.61	L, B
12	.64	.58	L, B
13	.52	.67	L
14	.78	.63	
15	.64	.56	B
16	.68	.38	B
17	.52	.51	L, B
18	.68	.63	L, B
19	.72	.70	L
20	.53	.59	L, B
21	.51	.25	L
22	.55	.49	B
23	.49	.41	L, B
24	.47	.55	L, B
25	.53	.43	L, B
26	.55	.53	L, B
27	.32	.57	L, B
28	.60	.48	L, B
29	.43	.47	L, B
30	.48	.35	L, B
31	.55	.54	B
32	.45	.28	L
33	.53	.32	L
34	.39	.42	L, B
35	.44	.40	L, B
36	.39	.40	L, B
37	.35	.42	L, B

Table 2 cont.

Item	Difficulty	Discrimination[b]	Selected Item[c]
38	.37	.39	L, B
39	.36	.52	L, B
40	.34	.48	L, B
41	.27	.24	L, B
42	.25	.34	L, B
43	.28	.37	L, B
44	.33	.41	L, B
45	.32	.27	L, B
46	.27	.29	L
47	.28	.48	L, B
48	.23	.40	L, B
49	.17	.32	L
50	.18	.38	L, B

[a] 31 of 40 items were common to both tests (78%)

[b] as measured by biserial correlations

[c] L = item selected using the Logist program
B = item selected using the Bical program

Table 3

Descriptive Statistics for the Item Parameter Estimates

N = 2500 K = 40

	One-Parameter Model	Three-Parameter Model		
	b_i[a]	a_i[b]	b_i	c_i
Mean	.01	1.26	.74	.22
Standard Deviation	.98	.37	.87	.14
Range	-2.36 - 1.85	.81 - 2.00	-2.17 - 2.13	.11 - .43

a
Item difficulties for the one- and three-parameter models range between $\pm$ three. Negative values indicate easy items and positive values indicate difficult items.

b
Item discrimination for the three-parameter model ranges between zero and two. This value is assumed to be constant in the one-parameter model.

Table 4

Descriptive Statistics for the two 40 Item Tests

	One-Parameter Model	Three-Parameter Model
Mean[a]	20.53	18.64
Standard deviation	6.09	6.85
Internal consistency	.83	.83
Standard Error of Measurement	2.51	2.82

[a] this difference was statistically significant, $p < .05$

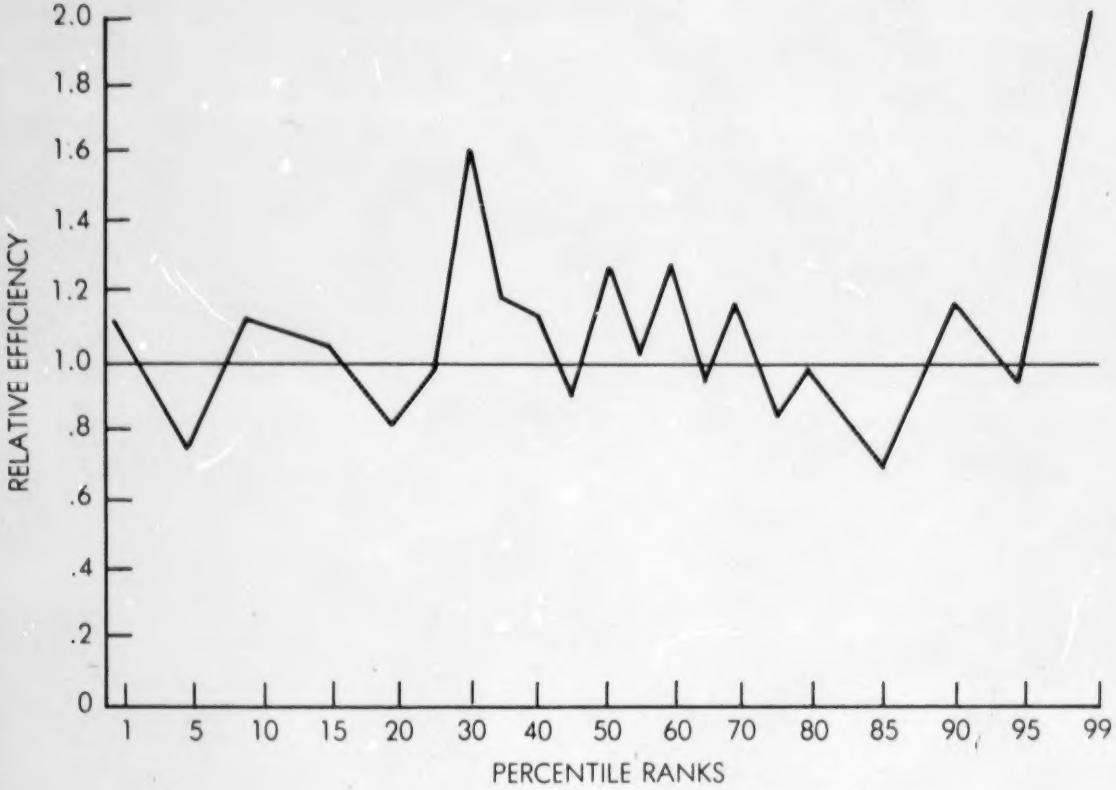

FIGURE 1. RELATIVE EFFICIENCY OF THE THREE-PARAMETER MODEL COMPARED TO THE ONE-PARAMETER MODEL.

Appendix A

Relative Efficiency values used in Figure 1

for the Comparisons between the two Latent Trait Models

Percentile Rank	Three-parameter model compared to one-parameter model
1	1.13
5	.76
10	1.14
15	1.06
20	.82
25	.99
30	1.61
35	1.19
40	1.14
45	.90
50	1.26
55	1.02
60	1.28
65	.96
70	1.17
75	.86
80	.98
85	.70
90	1.17
95	.97
99	2.06

References

Baker, F.B. Empirical comparison of the item parameters based on the logistic and normal functions. Psychometrika, 1961, 26, 235-246.

Benson, J. & Wilson, M. A comparison of three types of test development procedures using classical and latent trait methods. A paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Benson, J., Crocker, L. & Ware, W. A comparison of three types of item analysis in test development using classical and latent trait methods. A paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, Canada, March 1978.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Hambleton, R.K. & Cook L. Latent trait models and their use in analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Hambleton, R.K. & Traub, R.E. Information curves and efficiency of three logistic test models. British Journal of Mathematical and Statistical Psychology, 1971, 24, 273-281.

Hambleton, R.K. & Traub, R.E. Analysis of empirical data using two logistic latent trait models. British Journal of Mathematical and Statistical Psychology, 1973, 26, 195-211.

Hambleton, R., Swaminathan, H., Cook, L., Eignor, D. & Gifford, J. Developments in latent trait theory: A review of models, technical issues, and applications. A paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

Lawley, D.N. On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 1943, 61-A, 273-287.

Lazarsfeld, P.F. The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer et al. Measurement and Prediction. Princeton, N. J.: Princeton University Press, 1950.

Lord, F.M. A theory of test scores. Psychometric Monograph, 1952, No. 7.

Lord, F.M. The relation of test score to the trait underlying the test. Educational and Psychological Measurement. 1953, 13, 517-548.

Lord, F.M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lord, F.M. The relative efficiency of two tests as a function of ability level. Psychometrika, 1974, 39, 351-358 (a).

Lord, F.M. Quick estimates of the relative efficiency of two tests as a function of ability level. Journal of Educational Measurement, 1974, 11, 247-254 (b).

Lord, F.M. & Novick, M. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Mendenhall, W., Ott, L., & Schaffer, R. Elementary survey sampling. Belmont, Cal.: Wadsworth Publishing Co., 1971.

Rasch, G. Probablistic models for some intelligence and attainment tests. Copenhagen, Denmark: Denmarks Paedagogiske Institute, 1960.

Reckase, Mark D. Ability Estimation and item calibration using the one and three parameter logistic models: a comparative study. Research Report 77-1. Columbia Mo: Tailared Testing Research Laboratory, Educational Psychology Department, University of Missouri, 1977.

Wood, R.L., Wingersky, M.S., & Lord, F.M. Logist: A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, N.J.: Educational Testing Service, 1976.

Wright, B. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B.D. & Mead, R.J. CALFIT: Sample-free calibration with a Rasch measurement model. Research Memorandum No. 18, Statistical Laboratory, Department of Education, University of Chicago, 1975.

Wright, B.D. & Mead, R.J. BICAL: Calibrating rating scales with the Rasch model. Research Memorandum No. 23. Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1977.

Wright, B.D. & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.